DOCUMENT RESUME

ED 146 230 TM 006 633

AUTHOR Ross, G. Robert
TITLE A Factor Analytic Study of Inductive Reasoning Tests.
PUB DATE Apr 77
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Abstraction Tests; *Abstract Reasoning; Cognitive Tests; *Comparative Testing; Deductive Methods; Factor Analysis; *Factor Structure; High School Students; *Inductive Methods; Logical Thinking; *Measurement Techniques; Senior High Schools; Standardized Tests; Test Reliability; Test Validity; Verbal Learning; Visual Learning

ABSTRACT

A set of eight widely used inductive reasoning tests were investigated to determine whether or not they have different factorial structures. The eight inductive tests and three deductive tests, taken from the French Kit of Reference Tests for Cognitive Factors and the Watson-Glaser Critical Thinking Appraisal, were administered to 157 high school students. The reliability of each test was established, and factor analyses were performed on the intercorrelations. The results revealed that two different types of inductive reasoning were measured: verbally-based and perceptually-based. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED146230

D-39

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Robert Ross

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

A FACTOR ANALYTIC STUDY OF INDUCTIVE REASONING TESTS

G. Robert Ross, Ph.D.
Medical University of South Carolina

Presented to the Annual Meeting of the American Educational Research Association, April 1977

TM006 633

Abstract

Reasoning processes have traditionally been divided into two types, deductive and inductive. Although philosophers have been interested in inductive reasoning for centuries, psychological interest is a relatively recent occurrence beginning with the work of L. L. Thurstone. A review of psychological definitions of inductive reasoning revealed many diverse conceptualizations, several of which included deduction as a component of the inductive process. The number of tests which supposedly measure inductive reasoning were found to be equally numerous and diverse. The purpose of the subsequent study was to investigate the possibility that a set of eight widely-used inductive reasoning tests have different factorial structures.

The eight inductive tests and three deductive tests were administered to 157 high school students. The inductive and deductive tests were from the French Kit and the Watson-Glaser Critical Thinking Appraisal. The reliability of each test was determined and the intercorrelations were factor analyzed. Results indicated two types of inductive reasoning: verbally-based and perceptually-based. A major implication was that past research on inductive reasoning was instrument dependent.

Reasoning processes are fundamental to a large number of academic tasks at all levels of education. The learner is often required to cite a specific instance or exemplar of a general rule, or to generalize from a set of events or objects to a principle or rule that explains the set. Although the reasoning process has intrigued philosophers for centuries, and is obviously an important component of behavior required of school children, rigorous scientific investigation of this process is conspicuously lacking.

Traditionally, philosophers have dichotomized the reasoning process into deductive reasoning and inductive reasoning. Deductive reasoning is a process by which the nature of the premises of an argument can establish conclusively the validity or invalidity of the conclusion. An example may be seen in the following argument:

All A is B

This object is an A

Therefore, this object is a B

Here the validity of the conclusion is without doubt, given the premises stated. A deductive argument is valid when its premises, if true, provide conclusive grounds for its conclusion.

An inductive argument, however, is a process of examining the premises (or given information), and of generating a conclusion which is consistent with the premises. By the very nature of the process, the conclusion resulting from inductive reasoning can never be shown to be valid or invalid in the sense these terms are used with deductive reasoning. The argument following is an example of an inductive argument:

Socrates is a man and is mortal

Plato is a man and is mortal

Aristotle is a man and is mortal

Therefore, probably all men are mortal

(Copi, 1972, p. 24):

Note that there is always a probability associated with the conclusion from this type of inductive argument, a quality which is not present with a deductive argument.

In his formulation of a conceptual framework for inductive reasoning, Cox (Note 1) emphasized the point that although inductive reasoning is an important and wide-ranging process for knowledge generation, it is neglected in most educational curricula. Very few programs attempt to teach inductive processes. If it is to be included within future curricula, the process must be susceptible to reliable and standardized measurement in order for diverse research results to be generalizable to curriculum planning efforts.

This study is primarily concerned with the inductive reasoning process and the problems involved in measuring inductive behaviors. The research literature to date is encumbered with a lack of congruence in terminology, and has previously assumed equivalence in the wide range of psychometric instruments and laboratory tasks used to measure inductive reasoning. This paper will examine various attempts to measure induction and to what extent the laboratory and psychometric measures of induction are consistent among themselves.

Although philosophical interest in induction began with the writings of Aristotle, psychological interest in the inductive reasoning process has been a relatively recent occurrence. The first systematic study of inductive reasoning as a psychological phenomenon was conducted by L. L. Thurstone (1938), who labeled

induction as one of the eight "Primary Mental Abilities". Several subsequent studies also identified an inductive factor defined by various tests which required the subject to find a rule or principle which was consistent with the given information (Thurstone & Thurstone, 1941; Thurstone, 1944; Bechtoldt, 1947; French, 1951). Defined in this way, the process of induction may be considered a component of, or theoretically similar to, a number of different constructs in several research areas. Among these are the scientific method, artificial intelligence, intuition, concept formation, problem solving, and creativity.

Developmental studies of induction have been recently conducted by Fletcher and his associates (Cox & Fletcher, Note 2; Cox, 1972; Fletcher & Cox, 1973). These studies used tasks involving incomplete picture stimuli in which the subject was allowed to uncover segments of a familiar picture or pattern one at a time until a reasonable hypothesis was reported. In Cox and Fletcher (Note 2), the operational definition of induction was "utilizing information about some members of a class or scheme to support an inference about the whole or some larger portion of the whole" (p. 3), although this definition was shortened later to "specifying the whole pattern based on only partial information" (Fletcher & Cox, 1973, p. 1). For these studies, induction was equated with picture recognition. One criticism of this type of investigation is that it fails to control for repeated testing of unreported hypotheses, so that these investigators may simply be measuring deductive efficiency rather than induction. In any case, induction was apparently conceptualized as the ability to bring order to seemingly random stimuli.

Other researchers have attempted to define and measure inductive-like processes. Studies by Gollin (1960, 1961, 1962, 1965) and Smock (1955, 1957) primarily used tasks requiring subjects to identify a familiar pattern or picture

which is partially revealed. Westcott (1968) studied what he termed intuition; however, his experimental tasks were inductively based. The work of Guilford (1967) relating to divergent production and creativity involved administering tests which required inductive behaviors. Although preferring the term "concept formation," Bruner (1958) conducted many studies in which the subject was required to generate an explanatory hypothesis from incomplete stimuli. A great many investigators in the field of problem solving (see Davis, 1973) utilize tasks requiring subject behaviors which could be classified as inductive. Reviewing all these studies, Cox's (1972a) definition of induction perhaps best captures the psychological essence of the inductive process:

> Induction is the generation of a reportable hypothesis from only some members (as given) of a class, scheme, or pattern to describe the whole or at least some larger portion thereof. (The term hypothesis refers to a tentative explanation derived from insufficient evidence.) (p. 28)

In addition to the laboratory inductive tasks used by the investigators cited above, there have been a few attempts at a standardized measure of inductive reasoning. Probably the best known of these are the inductive reference tests contained in the Kit of Reference Tests for Cognitive Factors (French, Ekstrom, and Price, 1963). In collaboration with numerous investigators of cognitive abilities, French et al (1963) assembled over fifty tests which purport to measure 16 different cognitive factors, one of which is Induction. The French Kit inductive tests require the generation of hypotheses from pictures and drawings, as opposed to words or statements. In this respect they are very similar to (and in one case the same as) Thurstone's (1938) reference tests for induction.

Another standardized test which apparently requires inductive reasoning is the Watson-Glaser Critical Thinking Appraisal (Watson and Glaser, 1964).

Originally developed in connection with "fair-mindedness," the Appraisal contains five subscales, one of which - inference - requires inductive reasoning behaviors (as previously defined). Items on the Inference scale are in paragraph form, followed by several statements to which the examinee must assign a truth value. The five response alternatives for each statement are: True, Probably true, Insufficient data, Probably false, and False. The statements with "probable" alternatives represent behavior consistent with previous conceptualizations of induction, since an inductive generalization is inherently only probably true. Essentially the alternatives are hypotheses which may or may not follow from the given data.

The present problem with the measurement of inductive reasoning is the availability and use of too many untested instruments. Within any area of investigation researchers should attempt to base initial research results on a standardized measure of a given construct, or at least present evidence that newly-developed instruments are measuring the same construct. For example, there are vast superficial differences between the laboratory tasks used by Fletcher and Cox (1973) and those psychometric instruments used by Taba (1964), yet both claim to be measuring inductive reasoning. There has been no systematic effort to compare instrumentation, although the generalizability of research findings depends upon the assumption that the measurements are valid and reliable.

The problem, therefore, is that the measurement of inductive reasoning suffers from an apparent lack of construct validity, or at best, a lack of agreed upon definitions. Describing the problems associated with the measurement of psychological constructs, Nunnally (1967) states that:

> Any particular measure can be thought of as having a degree of construct validity depending on the extent to which results obtained from

> using the measure would be much the same if some other measure, or hypothetically, all the measures in the domain had been employed in the experiment. (p. 86)

With such a wide variety of inductive reasoning measures in use, it is questionable whether each measure obtains variance from the same construct. One of the purposes of this paper is to resolve this question empirically.

Measurement difficulties are not the only problems within the area of inductive reasoning, but they must be given first priority. A basic assumption here is that accurate and reliable measurement is the foundation for empirically-based constructs and theories. At the present time, researchers are "struggling to unscramble their concepts and clarify their language and models" (Wittgenstein, 1953, p. 21).

Naturally, the problem of construct validity is best approached empirically. To do this, Nunnally (1967) suggests a validation procedure based on the following steps: (a) specifying the domain of observables, (b) determining to what extent those observables correlate with each other or are affected alike by experimental treatments, and (c) determining whether or not all measures of such variables act as though they measure the construct (i.e., correlate in an expected way with measures of other constructs). These steps are rarely taken within any area of research, including inductive reasoning. Moreover, investigators have generally bypassed steps 1 and 2, proceeding to step 3 in hopes of finding interesting relationships between inductive reasoning and measures of other constructs. The overall motivation for this paper is expressed by Nunnally (1967) when he states "As the number of proposed measures of the same construct grow ... suspicion grows that they might not all measure the same thing". (p. 87)

This study attempted to investigate the reliability and factor structure of a representative set of nine widely-used inductive reasoning tests. Added to this set were tests of deductive reasoning which should theoretically be factorially distinct from the tests of induction. Initially, reliability coefficients obtained in this study were statistically compared to those reported by other authors. Given an acceptable reliability coefficeint and Kaiser's (1974) measure of sampling adequacy (MSA), scores on the tests of induction and deduction were subjected to a factor analysis. Two research hypotheses were of primary importance: 1) that the inductive and deductive tests would separate factorially if induction and deduction are different cognitive processes and 2) that the so-called inductive tests would form two factors related to test content (either verbal or pictorial). In addition, three experimental tests of induction were added to the analysis in an effort to determine their factorial structure. The overall purpose of the following study was to investigate what other authors have taken for granted, i.e., that all tests used in research on induction actually measure the same ability.

METHOD

Subjects

A total of 157 high school students were selected for participation in this study, including 77 (49%) and 80 (51%) from the 11th and 12th grades, respectively. Three students originally scheduled for testing were later unavailable due to outside commitments. All students attended a laboratory school associated with the Florida State University and were required to participate in the testing program. Subject ages ranged from 15 to 18, with 94% of the sample either 16 or 17 years of age. Among this group, there were 79 males and 78 females, with 117 white subjects and 40 black subjects. Standardized measures of intelligence indicated the mean IQ of subjects to be 105, with a standard deviation of 15.

For this study, subjects were not sampled randomly from a specified population for the following reasons: (a) all subjects were required to complete all reasoning tests at the Computer Applications Laboratory on the Florida State University campus, which precluded transporting students from elsewhere in Leon County, Florida, (b) the nature of the anticipated factor analytic treatment of the test scores demanded the maximum sample size possible (Harman, 1967), and (c) the laboratory school admits students on a stratified random basis from Leon County applicants. Therefore, the entire available number of subjects was utilized without further random sampling.

Apparatus

The apparatus for this study consisted of one test booklet per subject containing ten written tests. One additional test was administered via computer CRT terminal; subjects responded by pressing the appropriate keys. The eleven tests will hereafter be referred to collectively as the test battery.

The selection of tests to be included in the test battery was a process primarily influenced by the following considerations: (a) the primary purpose of this study was to investigate the possibility that the numerous instruments which purportedly measure inductive reasoning abilities do in fact measure different cognitive abilities, (b) all tests which supposedly measure inductive reasoning could not reasonably be included in the test battery, and (c) the test battery should include inductive instruments which represent both verbally and perceptually-based inductive tasks. Additionally, one of the themes of the literature review was that apparently several of the inductive tests reviewed required a substantial amount of deductive processing. Three tests of deductive reasoning, therefore, were also included to allow for a deductive reasoning factor, if one existed.

With these considerations in mind, a test battery was assembled containing a varied assortment of both inductive and deductive reasoning tests. Two hundred test booklets each containing ten written reasoning tests were printed by offset process and bound in a spiral booklet. The eleventh test (to be discussed in detail later) was developed by Dr. T. M. Wildman, Florida State University. A facsimile of the Inductive Matrices test is available from the author.

The total test battery consisted of eleven tests, comprised of (a) three inductive reasoning tests requiring substantial verbal skills, (b) two tests of perceptual closure which require inductive abilities according to Cox's (Note 1) definition, (c) three deductive reasoning tests, and (d) three experimental tests which have been used in previous research as measures of inductive reasoning ability. The first three tests were well-known tests of inductive reasoning from the Kit of Reference Tests for Cognitive Factors (French, Ekstrom, & Price, 1963). Specifically, these tests were (a) the Letter Sets test (I-1)[1], (b) the Locations test (I-2), and (c) the Figure Classifications test (I-3). According to the Manual for Kit of Reference Tests for Cognitive Factors, these tests

[1]French Kit code

measured "Associated abilities involved in the finding of general concepts that will fit sets of data, the forming and trying out of hypotheses" (French, Ekstrom, & Price, 1963, p. 19). They represent three content areas: letters, patterns, and numbers. In Guilford's (1967) terminology, two kinds of products were required: classes and systems.

It was intended that these three tests serve as marker variables for a verbally-based inductive factor, a procedure recommended by most investigators of cognitive abilities (see Cattell, 1966; Guilford, 1967). Naturally, the emergence of this expected factor, and of other factors defined by subsequent tests in the battery, was dependent upon the reliability of the obtained scores. In a previous paper, Stoker and Kropp (1967) reported test-retest reliability coefficients of .47, .48, and .58 for Tests 1, 2, and 3, respectively.

The next two tests in the test booklet, Tests 4 and 5, were selected for their apparent perceptually-based inductive component. They were also from the French Kit and were (a) the Gestalt Completion Test (Cs-1), a test utilizing incomplete pictures requiring subjects to mentally complete the picture and write its name in an adjacent space, and (b) the Concealed Words Test (Cs-2), a test closely resembling Thurstone's former Mutilated Words Test. In this test, the stimuli were words presented with parts of each letter missing and the subject responded by correctly identifying each word. According to the Kit Manual, these tests measure speed of closure, the ability to unify an apparently disparate perceptual field into a single percept. Stoker and Kropp (1967) reported test-retest reliabilities of .84 and .67 for tests 4 and 5, respectively.

Although not specifically called inductive reasoning tests, the rationale for including a set of perceptually-based instruments was that the responses required for these instruments appeared to be consistent with previously discussed definitions of inductive reasoning. Additionally, it was intended that

measured "Associated abilities involved in the finding of general concepts that will fit sets of data, the forming and trying out of hypotheses" (French, Ekstrom, & Price, 1963, p. 19). They represent three content areas: letters, patterns, and numbers. In Guilford's (1967) terminology, two kinds of products were required: classes and systems.

It was intended that these three tests serve as marker variables for a verbally-based inductive factor, a procedure recommended by most investigators of cognitive abilities (see Cattell, 1966; Guilford, 1967). Naturally, the emergence of this expected factor, and of other factors defined by subsequent tests in the battery, was dependent upon the reliability of the obtained scores. In a previous paper, Stoker and Kropp (1967) reported test-retest reliability coefficients of .47, .48, and .58 for Tests 1, 2, and 3, respectively.

The next two tests in the test booklet, Tests 4 and 5, were selected for their apparent perceptually-based inductive component. They were also from the French Kit and were (a) the Gestalt Completion Test (Cs-1), a test utilizing incomplete pictures requiring subjects to mentally complete the picture and write its name in an adjacent space, and (b) the Concealed Words Test (Cs-2), a test closely resembling Thurstone's former Mutilated Words Test. In this test, the stimuli were words presented with parts of each letter missing and the subject responded by correctly identifying each word. According to the Kit Manual, these tests measure speed of closure, the ability to unify an apparently disparate perceptual field into a single percept. Stoker and Kropp (1967) reported test-retest reliabilities of .84 and .67 for tests 4 and 5, respectively.

Although not specifically called inductive reasoning tests, the rationale for including a set of perceptually-based instruments was that the responses required for these instruments appeared to be consistent with previously discussed definitions of inductive reasoning. Additionally, it was intended that

these two tests would serve as marker variables for a perceptually-based inductive factor, if one existed. To this point, therefore, the test battery has been purposely constructed to contain two groups of tests which may act as marker variables in defining two separate inductive factors. Thus, the first five tests seek to confirm the factorial separation of two apparently diverse types of induction.

Tests 6, 7, and 8, on the other hand, were supposedly tests of deductive reasoning ability. Their inclusion in the test battery was based on the premise that the inductive measures previously reviewed required a substantial amount of deductive processing. Consequently, these three deductive tests should define a deductive reasoning factor which may or may not be related to scores on the inductive tests. Test 6, from the French Kit, was the Nonsense Syllogism Test (Rs-1), a test which required the subject to indicate whether the conclusion drawn from certain statements was correct or incorrect. Test 7, the Inference Test, was adapted for the French Kit from a similarly named test in Guilford's (1967) project. For each item, the subject was required to read one or more introductory sentences and to then select one of five conclusions that can be deductively drawn from these sentences.

The last deductive test, Test 8, was the Deduction test from the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1964). According to the Manual, this test measures "the ability to reason deductively from a set of given statements or premises; to recognize the relation of implication between propositions; to determine whether what may seem to be an implication or a necessary inference from given premises is indeed such" (Watson & Glaser, 1964, p. 2). The subject's task was to select those statements which deductively follow from a set of introductory statements and to circle the corresponding item number. Since subjects responded to all tests by marking in the test booklet, the test instructions were modified slightly to eliminate any reference to answer sheets.

Reported reliabilities for Tests 6 and 7 were .60 and .64, respectively (Stoker & Kropp, 1967), while the reliability for Test 8 was estimated to be .53, according to the test manual (Watson & Glaser, 1964, p. 14). Overall, the set of perceptually-based tests appear to yield the most reliable scores (average reliability = .76), followed by the three deductive reasoning tests (average reliability = .59) and the three inductive reasoning tests (average reliability = .51). If three distinct factors emerged from the initial eight variables, the perceptually-based tests were expected to have the greatest potential for defining an unequivocal factor due to their relatively high reliabilities.

In addition to the eight tests which comprise the core of the test battery, the inductive-deductive nature of three other tests were investigated. These tests were (a) the Inference test from the Watson-Glaser Critical Thinking Appraisal, (Test 9), (b) the Letter Series Completion test (Test 10) used by Simon & Kotovsky (1963), and (c) the Inductive Matrices test (Test 11) developed by Dr. T. M. Wildman, These three tests were labeled the "experimental tests" since very little was known about which abilities these tests measure. Test 9, the Inference test, measures "the ability to discriminate among degrees of truth or falsity of inferences drawn from given data" (Watson & Glaser, 1964, p. 2). The subject has five possible responses to each item: T(True), PT(Probably True), ID(Insufficient Data), PF(Probably False), or F(False). The items which require T and F responses are apparently deductive items, while the items requiring PT and PF responses are probably inductive. Again, the instructions to the student had to be modified to eliminate references to answer sheets. The Manual reports a reliability coefficient of .61 for this test.

The Letter Series Completion test, Test 10, required the subject to discover a pattern in a given series of letters and then to supply the next letter in the series. The test consisted of 15 items which no reported reliability coefficient.

The final test in the test battery, the Inductive Matrices test (Test 11), was considerably dissimilar in presentation and content from the other ten tests. First, the complete test was administered and scored by computer. The stimulus for this test was an 8 x 8 matrix, appearing on the computer terminal screen, consisting of the symbols X, O, and T arranged in some type of pattern (see Appendix A). The subject's task was to discover the pattern with as few errors as possible. Initially the pattern was completely concealed, after which the subject guessed at whether an X, O, or T was behind a particular concealed cell. Naturally, the first few responses were random guesses, but as the subject began to discover the pattern, his error rate dropped. The primary score of interest was the percent of correct responses per matrix. The complete test consisted of a series of four matrices, presented via computer CRT terminal, which took the average subject a total of approximately 40 minutes to complete.

In summary, the test battery consisted of 11 tests. The first eight tests were selected to provide marker variables for a three factor structure consisting of (a) a verbally-based inductive factor, (b) a perceptually-based inductive factor, and (c) a deductive factor. The remaining three tests were included primarily to determine their factorial composition, thereby providing needed evidence concerning the cognitive skills necessary for their completion. Table 1 below summarizes each test in terms of its name, number of items, and any reported reliability estimates.

(Table 1 about here)

Procedure

Subjects were divided into 16 groups of ten each. Groups 1 - 8 were eleventh graders and Groups 9 - 16 were twelfth graders. Two groups were tested daily spanning a period of eight school days, followed by an additional three days set aside for testing subjects who for various reasons could not participate at the scheduled time. The complete testing program extended over a period of approximately three weeks.

The testing process for each group of ten students was completed in one 3-hour session at the Computer Applications Laboratory (CAL) on the Florida State University campus. The CAL testing room accommodated ten students in a well-lighted, reasonably quiet, controlled climate condition. Each student was seated at a booth which prevented incidental conversation with his fellow students. The proctor introduced himself, handed out test booklets and pencils, and read a set of standardized instructions to the students.

In order to reduce the effects of fatigue in the latter stages of testing, four randomly generated orders of testing were utilized. That is, instead of completing each test sequentially in the test booklet, students were instructed to complete tests in the sequence indicated by the random ordering.

After completing six of the 11 tests, students were allowed a 10-15 minute break. Cokes were provided free and 99% of all subjects consumed at least 12 oz. of this caffeine-containing substance. Following the coke break, all groups completed the computer administered test (Test 11), lasting approximately 45 minutes. In the remaining time, students completed the last four tests in the test battery. Although the total testing process including the break lasted less than three hours, the total testing time was only two hours.

Results

The raw responses obtained from each test booklet were manually transferred to standard answer sheets and each test subsequently scored by an optical scanning device. The scoring formula used was that prescribed by the author and/or publisher. In addition to raw scores, internal consistency reliability coefficients were obtained for each test. The resulting basic data matrix, therefore, contained measurements on each of the eleven reasoning tests for each subject.

Difficulty and distributional characteristics

An examination of the difficulty of each test (defined as the maximum possible score ÷ the mean) revealed that the Locations test (T2) was the most difficult test overall (difficulty = .30). Test 11, the Inductive Matrices test, was the least difficult test (difficulty = .73). The majority of the other tests had a difficulty of around .50.

Since the difficulty of the test (and other factors) may cause the distribution of sample scores to deviate from normality, statistics related to the shape of the distribution of sample scores were also calculated. Measures of range, skewness and kurtosis for each variable did not differ substantially from those values expected from a normally distributed population; consequently, no transformation of the original data was made. Additionally, the assumption of interval data for all test scores was considered tenable, although this assumption is probably debatable.

Preliminary Tests

Ideally, the variance in observed test scores should be only a function of individual differences in ability for the most interpretable factor analytic results (Cattell, 1966). Preliminary tests for extraneous sources of variance indicated that differences in mean scores for each test were statistically

insignificant for groups defined by 1) testing order, 2) grade in school, 3) sex, 4) age, and 5) date tested. The effects of these unwanted sources of variance were, therefore, considered to be negligible.

Reliability and sampling adequacy

Another unwanted source of observed test score variance is that caused by random testing errors (error variance). Since the ability of factor analysis to display patterns of relationship in the data is partly an inverse function of the amount of measurement error contained in the scores (Harman, 1967), reliability estimates (1 - error variance) for each variable were examined. Also Kaiser's (1974) measure of sampling adequacy (MSA) is a useful index in determining whether the correlation matrix relating all variables is worth factoring. His present calibration of the index is as follows:

In the .90's - marvelous
In the .80's - meritorious
In the .70's - middling
In the .60's - mediocre
In the .50's - miserable
Below .50 - unacceptable

Reliability estimates obtained for tests T1 through T11 were alpha coefficients and are presented below in Table 2, along with the MSA value for each variable and for the total correlation matrix. The reliability estimates reported by either Stoker & Dropp (1967) or Watson & Glaser (1964) for tests T1 - T9 are presented for comparison purposes and statistically compared to the coefficients from the current study under the hypothesis of no significant difference, i.e., $H_o: \rho_1 - \rho_2 = 0$, with a two-tailed alternative. The inference concerning the difference in population correlation coefficients was made using the standard Fisher Z-transformation of r as outlined in Glass & Stanley (1970, p. 311). For tests T10 and T11, no reliability estimates existed for comparison.

(Table 2 about here)

It is evident from Table 2 that, in the majority of cases, the MSA was at least "meritorious". The MSA value of .82 for the complete R matrix indicated a psychometrically acceptable matrix for factorization. Also, the reliabilities obtained in the present study are greater than those reported elsewhere for all tests except one (Test 6). In two instances, for Tests 1 and 3, the differences in reliability coefficients are significantly different. The general trend toward higher reliabilities may partly be a function of the small-group testing format. Whatever the reasons for the difference, the reliabilities obtained in the present study are considered sufficient by most authorities not to preclude a factorization of the correlation matrix (Harman, 1967).

Factor Analysis

The first issue considered was the number of orthogonal common factors that would emerge from an analysis of tests T1 - T8. According to French, Ekstrom, and Price (1963), these tests should define three factors which may be called verbally-based induction, perceptually-based induction (or closure), and deduction. However, an examination of the logical content of the tests revealed a substantial amount of deductive processing required for the verbally-based inductive tests (T1 - T3). From this examination a hypothesis of two factors, general reasoning and perceptually-based induction, was postulated.

To investigate the factor structure of tests T1 - T8, the raw scores for all subjects (N) on each of the eight tests (n) were arranged in an (N x n) score matrix S. The raw data in S were then used to obtain all the Pearson product-moment correlation coefficients among the n tests, yielding a square correlation matrix, R_1, containing 1.0 as diagonal elements. These correlations appear in Table 3 below.

(Table 3 about here)

Since factor analysis, as opposed to principal components analysis, is concerned with extracting (explaining) common test variance rather than the total test variance, communality estimates were substituted for the diagonal elements in $\underline{R}_1$. The communality of a test (h_j^2) is defined as that part of the total variance the test shares with the common factors which emerge (Thurstone, 1947). Following common practice, the squared multiple correlation (SMC) was used as an initial estimate of communality for each test. Communality estimates were substituted into $\underline{R}_1$, resulting in the reduced correlation matrix, $\underline{R}$.

From the four different types of factor analysis currently in use (principal axis, canonical, alpha, and image factor analysis), the method of canonical factor analysis was selected as the most appropriate type. Its advantages included a test of factor significance (Rao, 1955) and the fact that the analysis is based on the assumption that a sample of subjects was chosen for testing. Additionally, in an extensive comparison of both the theoretical and practical aspects of the various methods, McDonald (1970) and Mulaik (1973) each concluded that the canonical analysis was the preferred type.

The problem of the number of common factors to extract has plagued factor analysis since its inception. The maximum number is naturally $\underline{n} = \underline{m}$, but this defeats the purpose of a supposedly reductionistic procedure. The problem, therefore, is to identify the optimum number of factors which can be extracted before the intrusion of non-common variance becomes serious (Cattell, 1966). Retaining only those factors with eigenvalues greater than 1.0 when all communality estimates are unity has been suggested by Kaiser (1959), but

there is a tendency for this method to extract a conservative number of factors. Cattell (1966) has proposed the "scree" test, which basically keeps all factors above the point at which the plot of eigenvalues versus number of factors becomes linear. Both methods are somewhat arbitrary and a subject of continuous debate in the literature. This study attempted to deal with the problem of the number of factors to extract from an eclectic position, i.e., several methods were utilized where appropriate, including Rao's statistical test.

The final choice to be made was the type of rotational procedure to be used after factorization. Since most previous studies of mental abilities have utilized an orthogonal rotation of factor reference axes, this method was used exclusively. To expect a domain of intellectual behavior such as reasoning to be explainable in terms of uncorrelated factors may seem unrealistic. However, for this study to be comparable to the previous factorial studies on reasoning abilities, keeping the same rotational procedure was essential. Furthermore, to allow the factors to be correlated (i.e., an oblique rotation) posed additional problems and arbitrary value judgments. To summarize, canonical factor analysis was always applied to $\underline{R}$, followed by orthogonal rotations of the factor references axes. The number of common factors was determined utilizing a variety of methods.

The reduced correlation matrix for tests T1 - T8 was factored using computer programs contained in the Statistical Package for the Social Sciences (Nie, Bent, Hull, 1970). Initially, the SPSS program FACTOR computes a principal component solution, i.e., factorization with 1.0 as diagonal elements and $\underline{m} = \underline{n}$. The eigenvalues and percent of total variance accounted for by each factor are one of several aids previously mentioned as indicators of the number of common factors present in $\underline{R}_1$. These values are presented in Table 4 below for the first five factors.

(Table 4 about here)

Using Kaiser's criterion of eigenvalues greater than 1.0 yielded two common factors, although the third factor closely approached this value. When Cattell's scree test was applied, it indicated three factors. Rao's (1955) test of factor significance also indicated that there were three factors. The four factor solution indicated only one substantial loading or the fourth factor after rotation; therefore, it was rejected as being inconsistent with the definition of a common factor (Thurstone, 1947). The two factor solution appeared to force all of the verbally-based reasoning tests, both inductive and deductive, to collapse into Factor 1. Furthermore, the two factor solutions did not agree with any of the criteria for determining common factors except Kaiser's criterion. From a statistical viewpoint, the Rao test of factor significance (Rao, 1955) indicated that the third factor was significant at the .05 level (χ^2 (13) = 22.8, p <.05). The three factor orthogonal solution, therefore, was considered to be the best overall description of the underlying common factor structure for the first eight tests. The rotated factor matrix is shown below in Table 5.

(Table 5 about here)

Once the number of common factors has been determined and the F matrix rotated, the variables which clearly define these factors must be examined to identify the nature of the factor. Technically, after orthogonal rotation each test vector has been resolved into a orthogonal component vectors of varying length. Those variables which have the larger components ("loadings"), on a particular factor axis tend to reveal the nature of the factor and imply a tentative label (Thurstone, 1947; Cattell, 1966).

An examination of the factor loadings for Factor I indicated that tests T1, T2, and T3 have their largest loadings on this factor, along with T7. Tests T1 - T3 were the verbally-based inductive tests from the French Kit, so this factor was tentatively labeled Verbally-based Induction (VBI). The high loading of T7 on Factor VBI was difficult to interpret since this test was supposedly a deductive test. For Factor II, tests T4 and T5 seemed to clearly define this factor and it was subsequently labeled Perceptually-based Induction (PBI). Factor III was most clearly defined by tests T6 and T8, both deductive tests. T7, and to some degree T8, contributed to Factor III but also to Factor I (VBI). Factor III was tentatively called a deductive factor (Factor D), but the factorial composition of T7 and T8 was apparently complex. These two tests contained both inductive and deductive components, a finding not predicted by a logical analysis of the content of the deductive tests. In subsequent analyses, the original three-factor structure was considered as the most meaningful standard for investigating the experimental tests.

Factorial Structure of the Experimental Tests

As indicated earlier, little was known about the three experimental tests T9, T10, and T11, except that they have each been used as measures of inductive reasoning in previous research. To recall, these tests were

T9 - The Inference Test from the Watson and Glaser Critical Thinking Appraisal

T10 - The Letter Series Completion Test used by Simon and Kotovsky (1963)

T11 - The Inductive Matrices Test developed by Dr. T. M. Wildman and administered via computer terminal.

The logical content of these tests has been reviewed in detail in elsewhere (Ross, Note 3). From this review, it was hypothesized that T9 and T10

would load highly on Factor VBI due to their apparent verbal inductive nature, although T9 appeared to require some deductive processes. Test 11, however, seemed to require more perceptual closure skills, much the same as Tests 4 and 5. Consequently, it was expected that T11 would load highly on Factor PBI.

Once an acceptable factor structure has been established, one of the most scientifically defensible methods for adding "new" tests was originated by Dwyer (1937). Dwyer's method allows the rotated axes of the original 3 - factor solution to remain stationary and then estimates the loadings of the new tests in that predefined factor space. In this way, the elements of the original (8 x 3) F matrix are unaltered. The three new tests, therefore, contribute nine new loadings (3 new tests x 3 old factors) at the foot of the original F matrix. The advantage of this procedure was that it allowed an explanation of the factorial structure of the experimental tests in terms of previously-defined factors. The formulae and further explanation are contained in Fruchter (1954).

The new rotated F matrix resulting from the application Dwyer's method is presented in Table 6 below.

(Table 6 about here)

Note that the factor loadings for tests T1 - T8 are equal to the original three-factor solution. Factor loadings for the experimental tests, T9 - T11, estimated by Dwyer's method, were added to the factor matrix. In this way, new loadings on the already-identified Factors VBI, PBI, and D can be more readily interpreted. That is, the factors need not be renamed; they have been identified previously.

The data in Table 6 revealed an interesting pattern of factor loadings for the experimental tests. Test 9 showed substantial loadings on both factors VBI and D with Dwyer's method. Test 10 consistently had large loadings on the VBI factor, indicating a strong verbally-based inductive component. Test 11, the test with the lowest communality (.24) but with good reliability (.75), tended to load well on Factor VBI. The hypothesis that T9 would have its largest loading on Factor VBI was confirmed, although a large deductive component was also evident. The hypothesis that T10 would load heavily on Factor VBI was also confirmed; in fact, T10 defines the Factor VBI better than any of the French Kit inductive tests. Contrary to its hypothesized structure, T11 did not load appreciably on Factor PBI, despite its apparent perceptual nature. The overall structure using Dwyer's method was consistent with an earlier description of the test content (Ross, Note 3), e.g., it was not surprising that Test 9 loaded highly on both an inductive and deductive factor since the items apparently called for both types of reasoning. The relatively poor loadings of Test 11 may be explained by a large uniqueness component in its observed score variance, indicated by its high reliability (.75) but low communality (.24). This result is not surprising either considering the unique nature of Test 11 as described earlier.

Discussion

The most important conclusion from this study was that two different types of inductive reasoning were identified in a representative sample of inductive tests: verbally-based induction and perceptually-based induction. Although the tests that measure these two types of induction were obviously different in content, no empirical evidence existed previously to support this difference. Unfortunately, research results which were based on either type test are only generalizable to that particular domain. In fact, evidence from the present study suggests that the VBI and PBI abilities are relatively unrelated. In future studies of induction, the type of induction being investigated should be emphasized by the author.

The question of which test best represents the VBI or PBI factor cannot be answered categorically since all tests in either domain were not studied. However, considering only the tests used in this study, some suggestions can be made. The test which loaded most heavily on the VBI factor was Thurstone's Letter Series Completion test; evidently this test would be most appropriate as a measure of verbally-based inductive skills. The Concealed Words test, on the other hand, was apparently the best measure of perceptually-based inductive skills. If only one test can be used, then the above mentioned tests are the logical choice. Further studies of factor structure within these domains may identify more refined tests which better represent their respective domains.

As a final note, the question of whether the three factors represented three mental abilities related to reasoning is debatable, although the results were reasonably consistent with a logical analysis of test content (Ross, Note 3). Naturally, the names of the factors are always debatable; it is conceivable

that all factors represented various types of deductive abilities. Moreover, it is possible that the inductive-deductive dichotomy is a poor way to conceptualize reasoning abilities. Perhaps the two are inseparable. These questions and others can be addressed by further research.

REFERENCE NOTES

1. Cox, W. F. Inductive Reasoning: A literature review and empirically oriented conceptualization. Unpublished manuscript, Florida State University, 1972.
2. Cox, W. F., & Fletcher, H. J. Children's inductive reasoning performances on picture stimuli. Unpublished manuscript, Florida State University, 1971.
3. Ross, G. R. Inductive reasoning: Conceptualizations, models, and measurement considerations. Unpublished manuscript, Florida State University, 1974.

REFERENCES

Bechtoldt, H. P. Factorial investigation of the perceptual speed factor. American Psychologist, 1947, 2, 304-305.

Bruner, J. S., Goodnow, J. J., and Austin, G. A. A study of thinking. New York: Wiley, 1958.

Cattell, R. B. The meaning and strategic use of factor analysis. In R. B. Cattell (Ed.), Handbook of Multivariate Experimental Psychology. Chicago: Rand McNally, 1966.

Copi, I. M. Introduction to logic (4th ed.). New York: Macmillan, 1972.

Cox, W. F. Developmental aspects of inductive and deductive reasoning. Unpublished doctoral dissertation, Florida State University, Tallahassee, 1972.

Davis, G. A. Psychology of problem solving. New York: Basic Books, 1973.

Dwyer, P. S. The determination of the factor loadings of a given test from the known factor loadings of other tests. Psychometrika, 1937, 2, 173-178.

Fletcher, H. J., & Cox, W. F. Children's inductive reasoning performance on incomplete picture stimuli. Working paper 13. Tallahassee, Fla.: The Florida State University Computer Applications Laboratory, 1973.

French, J. W. The description of aptitude and achievement tests in terms of rotated factors. Psychometric Monographs, 1951 (No. 5).

French, J. W., Ekstrom, R., & Price, L. Manual for the kit of reference tests for cognitive factors. Princeton, N.J.: Educational Testing Service, 1963.

Fruchter, B. F. Introduction to factor analysis. New York: Van Nostrand, 1954.

Glass, G. V., & Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Gollin, E. G. Developmental studies of visual recognition of incomplete objects. Perceptual and Motor Skills, 1960, 11, 289-298.

Gollin, E. S. Further studies of visual recognition of incomplete objects. Perceptual and Motor Skills, 1961, 13, 307-314.

Gollin, E. S. Factors affecting the visual recognition of incomplete objects: A comparative investigation of children and adults. Perceptual and Motor Skills, 1962, 15, 583-590.

Gollin, E. S. Perceptual learning of incomplete pictures. Perceptual and Motor Skills, 1965, 21, 439-445.

Guilford, J. P. The nature of human intelligence. New York: McGraw-Hill, 1967.

Harman, H. Modern factor analysis (2nd ed.). Chicago: University of Chicago Press, 1967.

Kaiser, H. F., & Dickman, K. W. Analytic determination of common factors. American Psychologist, 1959, 14, 425. (Abstract)

Kaiser, H. F., & Rice, J. Little jiffy mark IV. Educational and Psychological Measurement, 1974, in press.

McDonald, R. P. The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. British Journal of Mathematical and Statistical Psychology, 1970, 23, 1-21.

Mulaik, S. A. The foundations of factor analysis. New York: McGraw-Hill, 1972.

Nie, N. H., Bent, D. H., & Hull, C. H. Statistical package for the social sciences. New York: McGraw-Hill, 1970.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Rao, C. R. Estimation and test of significance in factor analysis. Psychometrika, 1955, 20, 93-111.

Simon, H. A., & Kotovsky, K. Human acquisition of concepts for sequential patterns. Psychological Review, 1963, 70, 534-546.

Smock, C. The influence of stress on the perception of incongruity. Journal of Abnormal and Social Psychology, 1955, 50, 354-356.

Smock, C. The relationship between "intolerance of ambiguity," generalization, and speed of closure. Child Development, 1957, 28, 27-36.

Stoker, H. W., & Kropp, R. P. Note on the kit of reference tests. Educational and Psychological Measurement, 1967, 27, 1171-1172.

Taba, H. Thinking in elementary school children. Cooperative Research Project No. 1574, San Francisco State College, 1964.

Thurstone, L. L. Primary mental abilities. Psychometric Monographs, 1938 (No. 1).

Thurstone, L. L. A factor analysis study of perception. Psychometric Monographs, 1944 (No. 4).

Thurstone, L. L. Multiple-factor analysis. Chicago: University of Chicago Press, 1947.

Thurstone, L. L., & Thurstone, T. G. Factorial studies of intelligence. Psychometric Monographs, 1941 (No. 2).

Watson, G., & Glaser, E. M. Manual for the Watson-Glaser Critical Thinking Appraisal. New York: Harcourt, Brace Jovanovich, 1964.

Westcott, M. R. Toward a contemporary psychology of intuition. New York: Holt, Rinehart, and Winston, 1968.

Wittgenstein, L. Philosophical investigations. Oxford: Blackwell, 1953.

Table 1

Summary Information on Tests Contained in the Test Battery

Test Code	Name of Test	Items	Reliability[a]
T1	Letter Sets test	15	.47
T2	Locations test	14	.48
T3	Figure Classifications test	14	.58
T4	Gestalt Completion test	20	.84
T5	Concealed Words test	25	.67
T6	Nonsense Syllogism test	15	.60
T7	Inference test	10	.64
T8	Deduction test	25	.53
T9	Inference test	20	.61
T10	Letter Series Completion test	15	Unknown
T11	Inductive Matrices test	4	Unknown

[a]See text for sources of reliability estimates.

Table 2

Measure of Sampling Adequacy (MSA) and

Reliability Estimates (α) for Tests

T1 Through T11

Test	MSA	α (Current Study)	α (Other Studies)	Fisher[a] z
T1	.83	.72	.47	3.69*
T2	.90	.58	.48	1.29
T3	.82	.95	.58	10.83*
T4	.73	.87	.84	1.04
T5	.65	.77	.67	1.94
T6	.78	.56	.60	-.56
T7	.80	.71	.64	1.19
T8	.89	.63	.53	1.40
T9	.84	.72	.61	1.84
T10	.85	.68	N/A	----
T11	.91	.75	N/A	----
(Total = .82)				

[a]An asterisk indicates the z statistic is significant at the .05 level.

Table 3

Correlations Between Tests T1

Through T8

	T1	T2	T3	T4	T5	T6	T7	T8
T1	1.0							
T2	.25	1.0						
T3	.33	.27	1.0					
T4	.20	.05	.17	1.0				
T5	.22	.06	.21	.42	1.0			
T6	.10	.21	.08	.20	.24	1.0		
T7	.35	.33	.44	.13	.06	.28	1.0	
T8	.28	.30	.29	.21	.18	.27	.57	1.0

Table 4

Eigenvalues and Percent of Total Variance for

Five Factors

(Principal Components Solution)

Factor	Eigenvalue	Percent of Variance
1	2.73	34.1
2	1.30	16.3
3	.98	12.3
4	.76	9.5
5	.67	8.4

Table 5

Varimax Rotated Factor Matrix

(3 Common Factors)

Test	Factor I	Factor II	Factor III	Communality (h^2)
T1	.52	.22	.09	.35
T2	.37	.00	.29	.24
T3	.63	.16	.07	.41
T4	.11	.59	.14	.40
T5	.11	.67	.07	.45
T6	.02	.26	.50	.33
T7	.59	-.03	.51	.57
T8	.42	.12	.56	.49

Table 6

Varimax Rotated Factor Matrix

Using Dwyer's Method[2]

Test	Factor I(VBI)	Factor II(PBI)	Factor III(D)	Communality (h^2)
T1	.52	.22	.09	.35
T2	.37	.00	.29	.24
T3	.63	.16	.07	.41
T4	.11	.59	.14	.40
T5	.11	.67	.07	.45
T6	.02	.26	.50	.33
T7	.59	-.03	.51	.57
T8	.42	.12	.56	.49
T9	.57	-.02	.53	.61
T10	.74	.27	.12	.63
T11	.42	.15	.20	.24

[2]Estimated loadings obtained by the Fortran program ESTLOAD, written by the author.